CREATING THE KNOWLEDGE ABOUT IT EVENTS

Gilad Barash, Ira Cohen, Eli Mordechai, Carl Staelin, Rafael Dakar
HP-Labs Israel
TRANSFORMING DATA TO KNOWLEDGE

Structured Data
- CRM data
- ERP data
- IT Measurements

Semi-structured Data
- System logs
- Events

Unstructured Data
- Forums
- Incidents
- Wikis
- Documentations

Sequences

Trees

Graphs
- UCMDB
- User links
EXAMPLE: DEBUGGING PROBLEM USING LOGS

03/15/2009 02:27 “Failed processing http request: report_ss_samples, from remoteHost :3.49.40.25 : Failed to acquire lock for publishing sample.”
TRANSFORMING EVENTS TO KNOWLEDGE

Semi-structured Data

Unstructured Data

• System logs
• Events
• Forums
• Incidents
• Wikis
• Documentations
COMPOSER /SEARCHER

Event

Composer
• Creates set of queries

Searcher
• Collects search results

“EJB spec violation Bean  Section 7.10.2 Warning A Session bean must implement directly”

EJB spec violation Bean  Section 7.10.2 Warning A Session bean must implement directly

EJB spec violation Bean  Section Warning A Session bean must implement directly

EJB spec violation Bean  Section Warning Session bean must implement directly

EJB spec violation Bean  Section Warning Session bean must implement
KNOWLEDGE CREATION SYSTEM

- Events
  - Creates set of queries
    - Composer
  - Collects search results
    - Searcher

- Quality of Information
- Associated Relevancy
- Source Rank

- Combines scores to create ranked
  - Ranker
- Knowledge Database
SOURCE RANKING

– To rank any source, s, we must solve the following set of equations to obtain the ranking u

$$\frac{\partial u}{\partial t} - f v + \nabla \cdot \nabla u = -\frac{\partial \phi}{\partial x} - \left( \frac{g \rho}{\rho_0} \right) \frac{\partial z}{\partial x} - g \frac{\partial \zeta}{\partial x} + \mathbf{F}_u - D_u$$

$$\frac{\partial v}{\partial t} - f u + \nabla \cdot \mathbf{v} = -\frac{\partial \phi}{\partial y} - \left( \frac{g \rho}{\rho_0} \right) \frac{\partial z}{\partial y} - g \frac{\partial \zeta}{\partial y} + \mathbf{F}_v - D_v$$

$$\frac{\partial T}{\partial t} + \nabla T = \mathbf{F}_T + D_T$$

$$\frac{\partial \mathbf{S}}{\partial t} + \nabla \cdot \mathbf{S} = \mathbf{F}_S + D_S$$

$$\mathbf{p} = \rho (T, \mathbf{S}, P)$$

$$\frac{\partial \psi}{\partial \mathbf{s}} = \left( \frac{-g H_z \rho}{\rho_0} \right)$$

$$\frac{\partial H_z}{\partial t} + \frac{\partial (H_z u)}{\partial x} + \frac{\partial (H_z v)}{\partial y} + \frac{\partial (H_z \Omega)}{\partial \mathbf{S}} = 0$$
SOURCE RANKING

Question: Which sources of documents (e.g., domains in www) are most relevant to the system I'm working on?

Method:

- Creates set of queries
  - Composer

- Collects search results
  - Searcher

- Extract domain names

- DomainScore += 1/rank

Repeat above for each event

Rank sources based on DomainScore

“EJB spec violation Bean Section 7.10.2 Warning A Session bean must implement directly”
**SOURCE RANKING: EXAMPLE RESULTS**

Networked printer logs
Multiple office HP laserjets. Logs collected from Microsoft event log.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Domain</th>
<th>Domain Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>hp.com</td>
<td>59.36915</td>
</tr>
<tr>
<td>2</td>
<td>microsoft.com</td>
<td>51.9311</td>
</tr>
<tr>
<td>3</td>
<td>eggheadcafe.com</td>
<td>36.79568</td>
</tr>
<tr>
<td>4</td>
<td>experts-exchange.com</td>
<td>33.69552</td>
</tr>
<tr>
<td>5</td>
<td>forums.techarena.in</td>
<td>25.29344</td>
</tr>
<tr>
<td>6</td>
<td>pcreview.co.uk</td>
<td>14.20567</td>
</tr>
<tr>
<td>7</td>
<td>tech-archive.net</td>
<td>14.05515</td>
</tr>
<tr>
<td>8</td>
<td>soft32.com</td>
<td>13.24757</td>
</tr>
</tbody>
</table>
KNOWLEDGE CREATION SYSTEM

Events

Composer
• Creates set of queries

Searcher
• Collects search results

Knowledge Database

Source Rank

Quality of Information

Associated Relevancy

Combines scores to create ranked

Ranker

Source Rank

Knowledge Database
QUALITY OF INFORMATION

– A measure of how fit the information is for a purpose

– Research Challenges:
  • Identifying important measures
  • Providing mechanisms to quantify and predict them
QUALITY OF INFORMATION FOR FORUMS

– Extract generic quality related measures for forums and incidents:
  • Ranking of users
  • Number of replies
  • Duration
  • ...

Challenge: Automatic methods for extraction from any forum type.

– Infer quality measures:
  • Was the question answered?
  • Which post(s) are answers / which are not
  • Difficulty of solution
  • ...

Challenges:
  • How to infer them? Can they be learned from other QOI measures?
PROCESS: INFER “ANSWERED/NOT ANSWERED”

Extract
• Collect forum threads
• Extract and compute generic features

Train
• Obtain labeled examples
• Train classifiers

Classify
• Use classifiers to label any forum thread
EXTRACT

- Java utility to download user forums and screen-scrape content elements

- Analyze and aggregate structured and unstructured features
PROCESS: INFER “ANSWERED/NOT ANSWERED”

Extract
- Collect forum threads
- Extract and compute generic features

Train
- Obtain labeled examples
- Train classifiers

Challenge: Label Noise
Users are responsible to change question from “not answered” to “answered”
LABEL NOISE – EXAMPLE

Using Log4j in WebSphere v6.1

This question is not answered.

Re: Using Log4j in WebSphere v6.1
Posted: Sep 24, 2008 04:31:22 PM
in response to: BSteinbach's post

BSteinbach,

Thank you very much. That works well.

Thanks again.
LABEL NOISE : THE PROBLEM

– Random label noise – does not occur around any class boundary

X – Class 1  
O – Class 2
SOLUTION: ENSEMBLE METHOD*

– Train N Classifiers with all training data

*Brody ET AL, journal of Artificial Intelligence research 1999
SOLUTION: ENSEMBLE METHOD

– Classify each sample with each classifier

Training Sample -> Classifier 1 -> Classifier 2 -> ... -> Classifier N

Ballot

Majority vote = Given label?

yes -> Add sample to new training data

no -> Discard training sample
SOLUTION 1: ENSEMBLE METHOD

– Train Classifier(s) with new training data

New training data

Classifier 1

Classifier 2

...

Classifier N
SOLUTION: ENSEMBLE METHOD + FLIP

– Classify each sample with each classifier

**Training Sample** → Classifier 1 → Classifier 2 → Classifier N → Ballot

- **Ballot**
  - **Majority vote = Given label?**
    - yes → Add sample to new training data
    - no → Randomly flip label based on classifier certainty, discard if not flipped
SOLUTION: ENSEMBLE METHOD + FLIP

- Train Classifier(s) with new training data
## NOISY LABELS: ACCURACY RESULTS

<table>
<thead>
<tr>
<th>Method</th>
<th>0%</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Noise Filter</td>
<td>0.78</td>
<td>0.75</td>
<td>0.73</td>
<td>0.69</td>
<td>0.65</td>
</tr>
<tr>
<td>Ensemble filter</td>
<td>0.78</td>
<td>0.77</td>
<td>0.75</td>
<td>0.72</td>
<td>0.69</td>
</tr>
<tr>
<td>Ensemble flip</td>
<td>0.78</td>
<td>0.77</td>
<td>0.75</td>
<td>0.73</td>
<td>0.70</td>
</tr>
</tbody>
</table>

*Results on UCI machine learning repository data*
PROCESS: INFER “ANSWERED/NOT ANSWERED”

Extract
- Collect forum threads
- Extract and compute generic features

Train
- Obtain labeled examples
- Train classifiers

Challenge: Transferability
Can a classifier trained on Forum A be used to classify threads on Forum B?
TRANSFERABILITY EXPERIMENT

Extract
- Collected 5500 Oracle forum threads, 1300 IBM forum threads
- Extracted 10 features

Train
- Training on threads from one domain, testing on the other

Classify

<table>
<thead>
<tr>
<th>Train/Test</th>
<th>Oracle</th>
<th>IBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>90%</td>
<td>85%</td>
</tr>
<tr>
<td>IBM</td>
<td>79%</td>
<td>97%</td>
</tr>
</tbody>
</table>
KNOWLEDGE CREATION SYSTEM

- Events: Creates set of queries
- Composer
- Collects search results
- Searcher
- Source Rank
- Quality of Information
- Associated Relevancy
- Combines scores to create ranked
- Ranker
- Knowledge Database
ASSOCIATED RELEVANCY

– Compute Levenshtein Distance between event and document
– Regular search engine may not have found the event but rather a collection of the words in the search string which are not related to each other
# PARIS SAMPLE RESULTS: HP ITRC FORUM

## Multifunction printers
- Product
- Print
- Scan
- Printer
- Multifunct
- Fax
- Copier

## NNM
- NNM
- Agent
- Insight
- Network
- Node
- Event
- OV
- Trap
- Monitor
- Alert
- SNMP
- SIM

## Databases
- Table
- SQL
- Connect
- Field
- Name
- Value
- Record
- DB

## HPUX
- HP
- HPUX
- UX
- Unix

## Proliant Servers
- MGMT
- Out
- Remot
- Pack
- Light
- Consol
- Pro
- Reset
- DL
- 380
- Liant
- Proliant
- ILO
- Firmware
- Lightsout
Status & Summary

• Created a system that gathered and reranked pertinent knowledge from the web to aid in troubleshooting and understanding system events in logs.
• System slated for HP Software’s BSM products
• Future work: Continue to refine feature selection and QOI measures